9-1-2012

Rawlsian Individuals: Justice, Experiments, and Complexity

John B. Davis
Marquette University, john.davis@marquette.edu

Published Version. *Journal of Economic Issues*, Vol. 46, No. 3 (September 2012): 729-743. Permalink.

JEI OURNAL OF ECONOMIC ISSUES
Vol. XLV I No. 3 September 2012
DOI 10.2753/JEI0021-3624460307

Rawlsian Individuals: Justice, Experiments and Complexity

John B. Davis

Abstract: John Rawls's *A Theory of Justice* is examined from the perspective of experimental methods in economics and complex adaptive systems simulations. This paper first discusses the justice principle selection process in Rawls's representation of it as a hypothetical experiment. This hypothetical experiment fails to satisfy reasonable experimental controls, particularly as reflects the conception of the individual it employs. The second section of the paper discusses the differences between Rawls's two conceptions of rational persons associated with his distinction between thin and full theories of the good. The third section uses his fuller conception of rational persons, life plans, and psychological laws in the third part of the book to offer an alternative view of the selection process understood as a complex adaptive system. The fourth section turns to a topic raised by this complex system approach, the status of normative reasoning in political-economic systems. The fifth section summarizes.

Keywords: complex adaptive systems, economic models, hypothetical experiments, normative reasoning, Rawls

JEL Classification Codes: D63, C9, B40, A12, A13

John Rawls's *A Theory of Justice* ([1971] 1999) is remarkably influential for a book that has arguably failed in a number of major respects.[1] Though powerful in its vision of a just society, and though receiving an incredible volume of commentary, Rawls's defense of its two principles of justice has hardly carried the day among philosophers, economists, and political scientists. Among other things, this implies that its method of argument, framed in terms of a hypothetical thought experiment designed to explain the principles of justice selection process, must ultimately be regarded as unsuccessful. Most of Rawls's critics, however, have rather focused on the substantive

John B. Davis is a professor of economics at Marquette University and professor of economics at the University of Amsterdam. *The author thanks Dave Colander, Steven Durlauf, Wade Hands, Harro Maas, Barkley Rosser, and two reviewers of this journal for helpful comments.*

normative principles in Rawls's thinking.[2] Few have examined his hypothetical experiment according to the standards of experimental science to determine whether it was likely to produce its expected outcome.[3] My focus is the latter, and in this paper I try to raise new issues about what we can learn from Rawls by asking questions that arise specifically for economists, particularly in light of relatively recent developments and new approaches in economics. The developments I am concerned with involve empirical methods of investigation associated with the laboratory and complex systems simulations. Viewing Rawls from these vantage points, I argue, casts a different light on his logic of moving from individuals in what he calls the original position to social outcomes, and also raises questions for economists regarding the status of the normative in economics.

The first section of the paper critically examines Rawls's main argument for his two principles of justice in terms of his representation of the selection process in the original position as a hypothetical experiment. The second section looks beyond the usual focus on Rawls's original position argument, and discusses the differences between his conception of rational persons in the first and third parts of the book associated with his distinction between thin and full theories of the good. The third section uses his fuller conception of rational persons in the third part of the book to offer an alternative view of the selection process as a complex adaptive system. The fourth section turns to a topic raised by this alternative view, namely, the status of normative reasoning in economics. The fifth section gives brief summary remarks.

The Original Position as if in the Laboratory

Rawls's famous two principles of justice are principles he argues individuals would choose for a society they would live in were they unable to determine what positions they would occupy in that society. The means he uses to explain this selection is a social contract idea he calls the original position that places individuals behind a veil of ignorance regarding most features about themselves that would be relevant to what places they might occupy in a society governed by any principles of justice. In the main argument of *A Theory of Justice* where this selection process is described (Chapter III, "The Original Position"), we are to imagine such individuals are presented with a "short list of traditional conceptions of justice" (including his own two principles), that they are "required to agree unanimously that one conception is best among those enumerated," and that their "decision is arrived at by making a series of comparisons in pairs" (Rawls [1971] 1999, 122-123). Five alternative conceptions of justice are listed with his own two principles placed at the top of the list (124). Rawls then imposes a number of conditions on the hypothetical selection process, including that the concepts of justice and right have been properly clarified for the individuals we are to imagine being involved (Rawls [1971] 1999, §§22 and §§23), that these individuals are properly behind the veil of ignorance in the sense that they must make their choices without regard to their future possible places (§24), and that the selectors behave as rational persons (§25). This set-up is summarized in twelve points that characterize the set-up for the selection process. Rawls then guides us through the

imaginary selection process itself by explaining how he believes individuals in the original position would reason their way through the alternative principles of justice (§§26-30). The result of this process is that his two principles are chosen by the imaginary selectors as principles of justice over the principle of average utility and classical utilitarianism as the main alternatives. Since the experiment is imaginary, Rawls himself acts as a surrogate for an imaginary selector by carrying out the reasoning process which he believes a representative individual would follow.

Rawls's original position idea is fully within the tradition of social contract thinking, but also involves a substantial innovation on that tradition in that it replaces the state of nature idea most social contract theorists have employed with the description of an imagined experiment framed in terms of scientific controls and proper experimental procedures. Individuals once behind the veil of ignorance, that is, are now seen as experimental subjects much as real experimental subjects in actual experiments who have had the rules of experiments explained to them. Further, Rawls judges the output of his imagined experiment in inductive terms as an experimentalist would, asserting that "no attempt is made to deal with the general problem of the best solution," and that he is accordingly limited by his framework to demonstrating the "weaker contention that the two principles would be *chosen* from the conceptions of justice on the . . . list" (Rawls [1971] 1999, 123; emphasis added). In effect, he implies, it is not his own deductive argument for the two principles of justice that carries the day but rather that of his imagined decision-makers in the controlled laboratory setting of the selection process.

> The decision of the persons in the original position hinges . . . on a balance of considerations. In this sense, there is an appeal to intuition at the basis of the theory of justice. Yet when everything is tallied up, it may be perfectly clear where the balance of reason lies. The relevant reasons may have been so factored and analyzed by the description of the original position that one conception of justice is distinctly preferable to the others. The argument for it is not strictly speaking a proof, not yet anyway; but, in [J.S.] Mill's phrase, it may present considerations of determining the intellect. (Rawls [1971] 1999, 124-125)

The two principles of justice, that is, are not determined by abstract reasoning, as was typically the case with Rawls's predecessors in the social contract tradition, but according to the intuitions of Rawls's imagined experimental subjects through their tallying up process guided ultimately by what they find preferable. This result may be recorded at most as telling us where "the balance of reason lies," so that following Mill, we learn what "considerations" may determine the intellect regarding justice. In all this, Rawls presents himself as standing outside of and supervising a selection process carried out by others, the results of which he of course favors, yet which his construction of the original position tells us is not produced by him but rather by his imagined experimental selectors.

This of course is a highly indirect way to establish a set of conclusions, and thus it is natural to ask why Rawls did not proceed in a more direct fashion by simply arguing for his two principles of justice using standard philosophical methods. He replies:

> Now admittedly this is an unsatisfactory way to proceed. It would be better if we could define necessary and sufficient conditions for a uniquely best conception of justice and then exhibit a conception that fulfilled these conditions. Eventually one may be able to do this. For the time being, however, I do not see how to avoid rough and ready methods. (Rawls [1971] 1999, 123)

There follows a discussion that would please experimental scientists regarding how this inductive process "singles out certain features of the basic structure as desirable," and a note of realism regarding the limitations of experimental practice in the assertion that "one cannot constructively characterize or enumerate all possible conceptions of justice, or describe the parties so that they are bound to think of them" (Rawls [1971] 1999, 123). That is, good scientific practice involves not over-reaching, showing caution about one's results, and allowing the evidence to generate valuable information about the considerations involved in people's thinking about justice. Rawls, then, rejects the more traditional deductive philosophical approach associated with defining "necessary and sufficient conditions for a uniquely best conception of justice" as an appropriate method of investigation, and presents his two principles as ones that would be discovered in a well-organized process of empirical observation.

Thus it seems fair to say that the reason Rawls adopts his indirect way of proceeding rather than argue directly for his conclusions is that he regards the procedure he uses as essentially more objective. Rather than engage in philosophical reasoning from his own perspective regarding the "necessary and sufficient conditions for a uniquely best conception of justice," and then have to attempt to "exhibit a conception that fulfilled these conditions" as he himself would personally see it, Rawls steps out of his role as philosopher, and allows what is "distinctly preferable to others" (Rawls [1971] 1999, 125) — others moreover who the veil of ignorance experimental controls makes into rational persons (§25) — to determine what the principles of justice are.

Of course, it comes as no surprise to the reader that these "others" select the two principles of justice which Rawls himself favors. Of course there are no "others" who actually make a selection among the principles of justice, since they are only imagined selectors, as Rawls's experiment is entirely a thought experiment. At the same time, in contrast to much of the long history of *Gedanken experimenten*,[4] his thought experiment is explicitly framed in modern laboratory terms as if it were being carried out in controlled experimental conditions. In fact Rawls appears sufficiently familiar with the problem of internal validity in laboratory experimentation that he explains the original position in terms of controls and procedural requirements in his

formulation of the veil of ignorance. Thus his original position idea innovates on the traditional thinking of the original contract theorists by treating the state of nature as if it were a carefully controlled laboratory, albeit an imagined one.

Ultimately, then, Rawls's reason for producing his results regarding the two principles of justice in this indirect manner, as opposed to direct philosophical argument, is that he takes the model of the laboratory to legitimate his results — results he apparently believes he cannot justify more directly as a single researcher. But the irony of his mode of presentation is that we all know that his experimental subjects are imaginary, and that there are difficult questions regarding his original position construction associated with the internal validity of his "experiment." Indeed, readers of *A Theory of Justice* cannot help but feel that the experimental set-up has been designed to weaken the attractiveness of positions Rawls opposes, as when he lists his own two principles of justice first, and poses as alternative justice principles ones he has long contested in the scholarly literature before the book's appearance. We might even speculate that since the book went through a lengthy gestation process and adjustment to many commentators (cf. the *Preface*), it is likely that Rawls's multiple drafts of the book involved his continually revising the set-up for the original position until he and others thought it gave as strong an argument as possible for his two principles of justice. In effect, Rawls considered possible outputs of previous runs of his "experiment," was not pleased with their results, re-worked the set-up, again examined the outputs, until, as a philosopher, he believed his arguments were consistent and persuasive. What the book ultimately does if we strip away the original position set-up, then, is what philosophers would expect it to do: make deductive arguments from well-examined assumptions. Yet the book is nonetheless rhetorically styled as an inductive procedure with independent experimental participants who produce Rawls's conclusions.

Many philosophers and other commentators have simply put aside Rawls's mode of presentation to concentrate directly on the relationship between the theoretical foundations he adopts in the original position and his two principles of justice. For example, some (e.g., Harsanyi 1975) have criticized Rawls's assumption that those selecting principles of justice from behind the veil would employ maximin reasoning (Rawls [1971] 1999, 152ff), since that plays a significant role in justifying the difference principle. The argument they make, however, is not that selectors might use other principles of reasoning that overturn the difference principle, but that it is not clear that *we*, reasoning philosophically about justice, should make maximin reasoning foundational to our theory of justice. From this perspective, Rawls's attempt to make it seem that maximin reasoning supports the difference principle, on the grounds that hypothetical experimental subjects in the original position could be expected to think in this manner, only gets in the way for the critics of asking whether individuals assumed to be free and equal (a characteristic of individuals in the state of nature) should indeed be conceived to use such reasoning. As a professional courtesy and measure of respect for Rawls, however, few say his experimental procedure is but a sleight of hand.

Suppose that economists nonetheless take seriously Rawls's idea that independent individuals might interact and select principles of justice. How might they otherwise proceed in explaining the selection or choice process? In the next two sections I suggest two changes to Rawls's "experimental" framework. First, we need a richer, more realistic conception of independent individuals as experimental subjects than he employs in the original position, and second we need a more realistic conception and modeling of the process of interaction between individuals than the one he employs. To address the first point, the next section turns to Rawls's often overlooked Part III (Chapter VII, "Goodness as Rationality") expanded discussion of what individuals are and his associated "full theory of the good" (Rawls [1971] 1999, 396). To address the second point, the following section uses this alternative conception of individuals and full theory of the good to model the justice principles selection process as a complex adaptive system.

Rawlsian Individuals and the Full Theory of the Good

We saw that an important part of Rawls's set-up involves placing "individuals" in his original position. That they are behind a veil of ignorance, but at the same time are able to choose principles of justice necessitates navigating a subtle balance between providing individuals sufficient abilities that they can judge prospective principles of justice and not providing them too much knowledge about how they would fare under different principles of justice. Rawls's strategy in Part I regarding how individuals are to be understood is to simply require that individuals are rational (Rawls 1971, §25). But by this he does not mean simply being capable of clear-headed reasoning, as one typically assumes for experimental subjects in social science in general. Rather, Rawls has a very specific sense of what is at play, since "the concept of rationality must be interpreted as far as possible in the narrow sense, standard in economic theory, of taking the most effective means to given ends" (Rawls [1971] 1999, 14). More fully:

> The concept of rationality invoked here . . . is the standard one familiar in social theory. Thus in the usual way, a rational person is thought to have a coherent set of preferences between the options open to him. He ranks these options according to how well they further his purposes; he follows the plan which will satisfy more of his desires than less, and which has the greater chance of being successfully executed. (Rawls [[1971] 1999, 143)

Indeed, a "feature of justice as fairness is to think of the parties in the initial situation as rational and mutually disinterested," where, if not strictly meaning they are egoistic, this means "they are conceived as not taking an interest in one another's interests" (Rawls [1971] 1999, 13). Thus in the opening argument of the book Rawls is working with at least a close analogue of the standard concept of a rational individual in economics, and views the selection or choice process across individuals as resulting in principles of justice by which individuals would order their lives together.

Yet once his two principles of justice have been selected, and after his Part II illustration of how those principles might function in modern constitutional democracies, Rawls proceeds to "present in more detail the theory of the good which has already been used to characterize primary goods and the interests of the persons in the original position" ([1971] 1999, 395) *en route* to his "full theory of the good" (396). Key to this full theory is an enlarged view of the individual built around the idea that "a person's good is determined by what is for him the most rational plan of life given reasonably favorable circumstances" (395; also cf. §15). What we find here, then, is not his earlier rational persons analogue to economics' rational individuals, but the development and application of the largely forgotten work of Josiah Royce (esp. Royce 1908), but also many other early twentieth century American writers — Ralph Barton Perry and John Dewey are emphasized (cf. Rawls [1971] 1999, 400n, 408n). Royce was influenced by Charles Peirce and William James, and as Rawls relates, advanced the idea that a "rational plan for a person determines his good," and that "a person may be regarded as a human life lived according to a plan" (408). There is much more that Rawls says about this life plan definition of the individual and also about what makes a rational plan of life rational, but for what follows I emphasize his following claim about a rational life plan's "time structure" (410):

> We must not imagine that a rational plan is a detailed blueprint for action stretching over the whole course of a life. It consists of a hierarchy of plans, the more specific subplans being filled in at the appropriate time. . . . The structure of a plan not only reflects the lack of specific information but it also mirrors a hierarchy of desires proceeding in similar fashion from the more to the less general. . . . A rational plan must, for example, allow for the primary goods, since otherwise no plan can succeed; but the particular form that the corresponding desires will take is usually unknown in advance and can wait for the occasion. (Rawls [1971] 1999, 410)

There then follow brief remarks on the subjects of scheduling and temporal sequence and also on the general idea that conflicting desires need to be avoided in order to maintain the person's more permanent aims and interests.

This is all very interesting and an under-appreciated dimension of Rawls's thought, but my goal here is not to fully present Rawls's enlarged view of the person, but to rather point to the departures it makes from the standard economics conception of rational individuals.[5] Essentially there seem to be three differences: (1) individuals understood to have rational life plans are heterogeneous because they encounter life's circumstances differently; (2) individuals do not have perfect foresight and are essentially boundedly rational; and (3) individuals engage in a deliberative kind of decision-making that involves weighing different sorts of considerations in making choices. Thus what motivates individuals at one point in their lives may not motivate them at another, and a life plan is not a "detailed blueprint for action stretching over the whole course of a life." Yet as being identified as "rational" life plans, individuals still always try to do the "best" for themselves. This is accordingly

where Rawls's expanded theory of the good goes beyond his schematic primary goods approach.

The underlying idea, moreover, is that people can always be seen as seeking to improve their capacities and abilities — a notion which he terms "the Aristotelian Principle":

> The Aristotelian Principle runs as follows: other things equal, human beings enjoy the exercise of their realized capacities (their innate or trained abilities), and this enjoyment increases the more the capacity is realized, or the greater its complexity. . . . The intuitive idea here is that human beings take more pleasure in doing something as they become more proficient at it, and of two activities they do equally well, they prefer the one calling on a larger repertoire of more intricate and subtle discriminations. (Rawls [1971] 1999, 426)

Thus individuals undergo a kind of personal development over their lives, which is guided by the fundamental principle of having a rational life plan, is affected by the twists and turns along life's various pathways, and varies from one individual to the next. While this remains a very general set of ideas about what individuals are, the point I want to make is that it more closely reflects thinking about individuals/agents in complexity economics examinations of social interaction than in standard economics' atomistic rational individual conception. Individuals as described in complex adaptive systems have a general motivation to do the "best" for themselves, but what this involves both varies across individuals — they are heterogeneous — and is affected by the pathways they follow in interaction with each other. Consider, then, the modeling of individuals thus understood in terms of their social interactions, and how this allows us to re-characterize the principles of justice selection process.

The Selection Process as Social Interaction in a Complex Adaptive System

How might we model a principles of justice selection process in terms of social interaction in a complex adaptive system? Rawls models his original position selection process as if it were a laboratory experiment, imagining that real individuals might participate as experimental subjects. But there is considerable ambiguity surrounding how this imaginary process might work itself out were Rawls not to pre-determine the outcome. In contrast, modeling the selection process as a complex adaptive system involves a more highly disciplined procedure. Complex systems are typically simulated because their properties cannot be determined analytically. In this respect, they are like Rawls's imaginary original position experiment in that they call for an inductive type of investigation. One has to run the simulation to discover what the properties of the system are. But in contrast to Rawls's method, using a complex system approach first requires identifying some mathematical-logical model of the process being investigated, which is then programmed for the simulation. There is really no counterpart to this in Rawls's investigation. Of course a variety of theoretical

assumptions and arguments play a role in his representation of the original position, but they lack the character of a model, at least as models are understood in economics. A simulation approach, then, is like an experimental one, but is different in being structured by its underlying mathematical-logical model.

Non-economists (and many economists too) often express skepticism regarding the scientific value of mathematical models, arguing they are too detached from the reality being modeled, and this might be thought to argue against simulating complex systems as an alternative means of investigating Rawls's selection process. Against this is the fact that beginning in the 1930s and especially in the postwar period mathematical modeling has been central to economic practice as specifically a method of empirical observation (cf. Morgan 2003). That is, in contrast to theories, mathematical models permit "numerical representation of the phenomena under investigation," which constitutes the means to their measurement, which is itself "a kind of observation" (Boumans 2005, 2). Without measurement, the quantitative dimensions of phenomena, particularly economic phenomena, are unobservable, and empirical science cannot go forward. What critics typically overlook regarding the modeling process is that economic reality is not there to be observed until it has been represented quantitatively. Formal modeling (plus the attendant programming for simulations), then, provides the foundations for empirical investigation of social systems made up of interdependent individuals that are too complex to investigate in more standard ways. Thus the argument for replacing Rawls's method with a complex systems approach is that the empirical investigation he wants to carry out necessitates a more sophisticated scientific apparatus than his *Gedanken experimenten* offer.

What, then, is involved in modeling the selection process as a complex adaptive system? That individuals are interdependent in complex adaptive systems means that their behavior needs to be interpreted as a stochastic process that represents the optimal behavior of each individual as conditional on the behavior of all others and also on the aggregate properties of the system they occupy. Individuals who are boundedly rational try to do the "best" they can for themselves, but what this specifically amounts to cannot be fully determined apart from the pathways they pursue, and these pathways themselves depend upon their interaction and the character of the system in which they interact. Such systems are described as complex and adaptive because feedback effects on individual behavior from interaction between individuals constantly causes individuals to revise (and we might say, "deliberate over") what they regard as the best thing to do. These behavioral adjustments in turn change the aggregate characteristics of the system, so that complex adaptive systems exhibit phase transitions and emergent properties that then further affect individual behavior. Complex systems may also be characterized by network externalities and sensitivity to shocks and random factors, which can also influence the sequencing of individual decision-making, and produce path dependence for individual behavior and the system as a whole (cf. Durlauf 2005).

Suppose, then, that individuals are understood as in Rawls's Chapter VII account of rational persons rather than as in his Chapter III rational economic individuals, and that the justice selection process is seen as working as a complex

adaptive system rather than as in his original position scenario. That is, we abandon Rawls's artificial and narrowly constructed imaginary experiment described above, and represent the principles of justice selection process as a process of social interaction understood in terms of a sequence of decentralized, uncoordinated decisions made by heterogeneous individuals, where the environment has a potential for network externalities and unexpected shocks. How do individuals behave? Rawls's Chapter VII view of individuals, whose rational life plans consist of "a hierarchy of plans, the more specific subplans being filled in at the appropriate time" (Rawls [1971] 1999, 410), suggests that individuals operate with a changing collection of different strategies, algorithms, or mental models, hierarchically arranged "from the more to the less general" (410) which they draw upon and continually revise in order to address the different types of circumstances they encounter. As some of these strategies are more successful than others, individuals settle on smaller sets of strategies that tend to best achieve their broad life plans (though this smaller collection of strategies is presumably still always subject to revision). That is, they engage in a learning process that generally improves their capacity to achieve broad plans of action that progressively promote their capacities and abilities. These broad plans are emergent upon this learning process, and again following Rawls they may be generally understood according to his Aristotelian principle whereby "human beings take more pleasure in doing something as they become more proficient at it" (426). So Rawls's broader conception of the person in terms of continually evolving life plans fits quite well with the view of individuals interacting with one another in complex adaptive systems.

Here I do not offer a particular model or simulation of such an approach, but instead make a general claim regarding what this alternative set-up tells us about a principles of justice selection process. Since complex adaptive systems are characterized as exhibiting emergent properties, it follows that we cannot guarantee that complex processes of social interaction between individuals will produce Rawls's two principles of justice *unless* we have a special reason to say that the learning process in which individuals are engaged somehow enables them to sort through their alternative strategies in such a way that they generally tend to settle on certain shared principles of interaction — which we might then call principles of justice. That is, something needs to ground the learning process in certain principles if Rawls's view is to go through. Here, then, we might note a set of suggestions specific to the nature of social interaction which Rawls makes following his discussion of life plans regarding what he believes lends "relative stability" to systems of justice. Thus in Chapter 75 he goes on from his discussion of life plans to state that there are three psychological laws which appear to underlie much human behavior: in families people return love, in communities they reciprocate fellow feeling, and institutions seen to be just engender a sense of just behavior.[6] If these principles indeed hold, then it seems reasonable to say his principles of justice could emerge from a complex interactive selection process. That is to say, should our specification of the behavioral properties of individuals modeled as interacting in complex systems make psychological characteristics such as these a part of their make-up, then, depending on such things as the conditions under

which they interact, Rawls's principles could be selected. This, of course, makes Rawls's account more complicated and a complex systems interpretation of the selection process is clearly a departure from his methodology, but it has the virtue of showing a way of linking substantive normative views more tightly to his broader view of the individual in terms of life plans and deep psychology. I draw three general conclusions, therefore, from this alternative representation of a justice principles "selection" process in order to comment on Rawls's project.

First, Rawls's innovation on the classic state of nature social contact idea in the form of his original position thought experiment seems to undermine rather than support philosophy's traditional analytic form of investigation of justice principles. Philosophy's comparative advantage as a discipline, we might say, is to engage in abstract reasoning regarding the most general categories of reality. Rawls, however, repackages this reasoning as a historical experiment in a choice process, and invests legitimacy in the conclusions reached only by virtue of this repackaging. In doing so he arguably jeopardizes the meaningfulness of the traditional abstract reasoning process, since a more realistic account of social choice processes, as I hope to have suggested in the alternative proposal set forth here, tells us that what can be achieved in this manner is more limited than Rawls believes. Indeed, much of the critical review of *A Theory of Justice* by philosophers concentrates on the abstract argument running from Rawls's characterization of individuals to his principles of justice, setting aside his misleading device of the original position and veil of ignorance. Better, it might be said, to have left philosophy as philosophy, and not to have moved onto the unstable terrain of economics' rational individual social choice analysis, unless one is prepared to think more systematically about the nature of social interaction.

Second, we might register an important lesson from complexity theory regarding aggregation procedures and social outcomes by noting that Rawls's device of replacing multiple individuals with a hypothetical representative individual in the justice principles selection process (Rawls [1971] 1999, 96ff) closely resembles Arrow's "dictator" solution to the problem of social choice (Arrow [1951] 1963), though without a recognition on Rawls's part of it as an impossibility result. Arrow's impossibility problem was that one cannot aggregate over the choices of multiple individuals (under reasonable assumptions) to produce a unique social choice function except when one sets aside the aggregation exercise by substituting the preferences of a single "dictator" for the preferences of all individuals (thus essentially rendering the social choice idea meaningless). Arrow's problem was subsequently shown to be species of a more general difficulty associated with aggregating up to social-level phenomena in individualist frameworks, for example, in connection with discovery of the problematic character of general equilibrium theory where aggregating up to system properties from properties ascribed to individuals — the so-called Sonnenschein-Mantel-Debreu results. More generally, it is now clear that it is often the case that claims made about either individuals or aggregates of individuals cannot be readily extended to the other without very strong assumptions (Kirman

2011, 84-86). Recognition of this entire issue and the nature of aggregation problems is unfortunately missing from Rawls's project as well as much of the commentary on it.

Third, we might re-appraise Rawls's results from the perspective of Part II of *A Theory of Justice*, where as noted already, he addresses the fit, so to speak, of his two principles selected in the first part of the book with the practices and history of constitutional democracies in the postwar period. Suppose for the moment that historical societies in the development of liberal market systems indeed work much like complex adaptive systems, an argument which has famously been made in a smaller space with respect to the technology adoption process in these societies (David 1985). Then we might argue that Rawls's statement and argument for his two principles of justice simply record in reflexive fashion a social outcome of the postwar period as observed by one person, John Rawls. His particular representation of the determination of these principles in terms of his Part I original position selection process then functions as an adequate myth in the tradition of state of nature arguments in Western liberal states, but his identification of his two principles of justice rather constitutes one observation of the justice principles (among many) that are in fact emergent in liberal market societies. On this view, Rawls's principles of justice (and also the rival principles of justice in play today) are social outcomes of interaction between individuals in relatively decentralized modern social systems. Thus Rawls's desired results might be achieved in a more realistic manner by historical argument, and the real strength of *A Theory of Justice* is then his Part II demonstration that his principles are operative in successful constitutional democracies, all on the assumption that the social-economic world works much as a complex adaptive system.

Some Comments on the Normative in Economics

What my arguments in the previous sections were meant to do was to give a reading of Rawls's *A Theory of Justice* from the perspective of recent economics, given that Rawls employs a method of argument that resembles new methodologies employed in recent economics. Since Rawls is investigating a choice process, and investigating the conditions under which choices get made has become increasingly central to recent economics, it is interesting to re-examine Rawls's from the point of view of the latter. While this provides new perspectives on how principles of justice might emerge from social interaction, there is a problem in proceeding this way when it comes to how we think about the normative. Whereas Rawls's book is explicitly a normative investigation into how his principles of justice might be defended as fair, economics, at least in mainstream theory, quite strongly separates descriptive and normative judgments. Indeed, the characterization of social interaction in complex adaptive systems as outlined in the last section also proceeds as if normative concerns are not really involved, so that by recasting Rawls in this manner we are in a position of having to say that should social interaction in complex systems generate Rawls's types of principles, then we *might* "call" them principles of justice. Is there a way, then, of

arguing that an interactive process that reframes Rawls in this way actually produces genuine principles of justice?

One way to answer this question consistent with complex adaptive systems theory, which also uses Rawls's richer life plans view of the individual with its emphasis on individuals' progressive improvement of their capacities and abilities, is to make recourse to the concept of a convention. Systems of social interaction often cause conventions to materialize, because they simplify decision-making. But conventions *per se* are only normative in a weak sense in that they are essentially pragmatic, and make no special reference to ethical values. It is interesting, then, that there exist experimental grounds for arguing that conventions sometimes evolve into social norms in a stronger normative sense. Francesco Guala (2010) recently did an experiment in the form of a coordination game that ran for multiple rounds, thus in some respects approximating a number of the features of a complex adaptive system whereby individuals in interaction with one another would learn which strategies worked best. He found that conventions about play emerged by the ninth round. Then in a tenth round he allowed one of the players to be tempted to violate this convention. The remaining players, however, continued to observe the convention, and he argued that this suggests that "conventions acquire normative power" as more than just pragmatic requirements (2010, 755). Why might we follow him in this conclusion? Here Rawls's life plan idea can be helpful, since we could say that conventions get integrated into individuals' life plans on account of their relative stability, and since these life plans are built around individuals' progressive realization of the good as individuals see it for themselves, these conventions are thereby elevated to social norms which individuals believe they ought to observe for moral reasons. Thus reading Rawls in terms of a complex social process of social interaction, where we replace his standard economics conception of rationality and the individual by his broader views of Part III, has the effect of not just recasting his view but also that of complex social systems by showing how they may come to exhibit normative properties.

Note again, then, that part of Rawls's thinking about individuals and their life plans is that he believes they tend to observe the three psychological laws distinguished above. Rawls associates these laws with the "relative stability" of systems of justice, but it is important to be clear about the direction of causality. Rawls's three laws can be seen to be principles of fairness since whether in the family, community, or society's institutions, the idea is that what one side gives is reciprocated, where this is seen as a matter of being fair. Thus it is not that relatively stable patterns of interaction, or conventions in Guala's experiment, evolve into social norms, but rather than there are norms implicit in systems of social interaction that behaviorally speaking are normative by virtue of the meaning of reciprocity. It is not, consequently, that systems of social interaction have the normative as an emergent dimension, but rather that the ways in which they are normative is emergent upon how people interact. This is relevant to whether the principles of justice in a process seen as complex and adaptive ought to be expected to turn out as Rawls believed. What he

might better have argued would happen is not that necessarily his two principles would be selected, but that principles would be selected that were stable and fair.

Summary Remarks

This paper takes another look at Rawls's *A Theory of Justice*, drawing on recent thinking in economics to re-appraise his method of investigation. Much of the early excitement about the book was due to his innovative framing of the justice principles selection process, but from the vantage point of recent economics this innovation is neither very remarkable nor very well done. In the first place, Rawls's imagined experiment falls well short of experimental practice in economics and social science in general, so much so that he more exposes his preferred principles of justice to question than gives them the credibility that a more direct form of argument would likely produce. Secondly, by making principles of justice the outcome of a social process he invites us to ask how and whether social processes generate such principles. Then, if we model social processes as complex and adaptive, we come up with quite different views about social outcomes and justices principles, and moreover find ourselves with many new problems regarding their status and the ways in which they evolve as normative principles.

Consequently it could be argued that what is most interesting about *A Theory of Justice* is not its arguments and the book itself, but rather the agenda it opens up for the theory of justice. By Rawls's lights, the developments in recent economics briefly described here raise questions that are prior to and in some respects more fundamental than his own investigation. How can we investigate characteristically normative selection principle processes in social processes? How do social systems generate and modify normative principles, especially ones that are over-arching and systemic in nature? Recent work in economics seems to be increasingly moving in the direction of these issues, but it remains to be seen whether it will take up Rawls's deep concerns, the nature of justice, and make them central to its development.

Notes

1. Citations from the text are to the original edition.
2. See Pogge (2007) for a comprehensive recent evaluation and review of Rawls's justice thinking. Rawls himself shifted his focus and method in his subsequent *Political Liberalism* ([1993] 1996).
3. See Frohlich, Oppenheimer and Eavey (1987a, 1987b) who actually implement Rawls's hypothetical experiment in the laboratory, and find that his principles of justice are not confirmed.
4. The expression was coined by Ernst Mach ([1905] 1975).
5. Note also that Rawls's life plan concept is an ideal one. As one reviewer pointed out, many people, for example the working poor and those below the poverty line, live payday to payday (see Wilson 1987).
6. I thank a reviewer from this journal for this reference.

References

Arrow, Kenneth. *Social Choice and Individual Values*, 2d ed. New Haven: Yale University Press, [1951] 1963.
Boumans, Marcel. *How Economists Model the World into Numbers.* London: Routledge, 2005.

David, Paul. "Clio and the Economics of QWERTY." *American Economic Review* 75 (1985): 332-337.

Durlauf, Steven. "Complexity and Empirical Economics." *Economic Journal* 115, 504 (2005): F225-F243.

Frohlich, Norman, Joe A. Oppenheimer and Cheryl L. Eavey. "Laboratory Results on Rawls's Distributive Justice." *British Journal of Political Science* 17 (1987a): 1-21.

——. "Choices of Principles of Distributive Justice in Experimental Groups." *American Journal of Political Science* 31 (1987b): 606-636.

Guala, Francesco. "How History and Conventions Create Norms: An Experimental Study." *Journal of Economic Psychology* 31 (2010): 749-756.

Harsanyi, John. "Can the Maximin Principle Serve as the Basis for Morality? A Critique of John Rawls's Theory." *American Political Science Review* 69 (1975): 594-606.

——. *Complex Economics: Individual and Collective Rationality*. London: Routledge, 2011.

Kirman, Alan. *Complex Economics: Individual and Collective Rationality.* London: Routledge, 2011.

Mach, Ernst. "On Thought Experiments." In *Knowledge and Error: Sketches on the Psychology of Inquiry*, translated by Thomas J. McCormack and Paul Foulkes. Boston: Dordrecht, [1905] 1975.

Morgan, Mary. "Economics." In *The Cambridge History of Science, vol. 7, The Modern Social Sciences*, edited by Ted Porter and Dorothy Ross. Cambridge: Cambridge University Press, 2003.

Pogge, Thomas. *John Rawls: His Life and Theory of Justice*, translated by Michelle Kosche. Oxford: Oxford University Press, 2007.

Rawls, John. *A Theory of Justice.* Cambridge: Harvard University Press, [1971] 1999.

——. *Political Liberalism.* New York: Columbia University Press, [1993] 1996.

Royce, Josiah. *The Philosophy of Loyalty.* New York: Macmillan, 1908.

Wilson, William Julius. *The Truly Disadvantaged.* Chicago: University of Chicago Press, 1987.